

The Ethical Landscape of Robotics

Paweł Lichocki, Peter H. Kahn Jr., Aude Billard

Cite as: P. Lichocki, P. Kahn Jr, and A. Billard. The Ethical Landscape of Robotics.
IEEE Robotics and Automation Magazine, 18(1):39-50, 2011.

Abstract—This paper surveys some of the main ethical issues pertaining to robotics that have been discussed in the literature. We start with the idea of responsibility ascription that arises when an autonomous system malfunctions or harms people. Next, we discuss various ethical issues emerging in two sets of robotic applications: service robots that peacefully interact with humans and lethal robots created to fight in battlefields. Then, we provide a short overview of machine ethics, a new research trend that aims at designing and implementing artificial systems with “morally” acceptable behavior. We then highlight resulting gaps in legislation, and discuss the need for guidelines to regulate the creation and deployment of such autonomous systems. Often, when designing such systems, the benefits tend to overshadow partly unknown but potentially large negative consequences.

Key words: Roboethics, Machine Ethics, Human-Robot Interaction, Autonomous Systems

I. INTRODUCTION

This decade has undergone a “robotic demographic explosion.” The number of industrial robots in operation exceeded 1 million at the end of 2008. Sales of robots for personal and domestic purposes have increased rapidly since 2000 and reached 7.2 million at the end of 2009 [41]. The rampant growth of service robots has led people to rethink the role of robots within society. Robots are no longer “slave” machines that respond only to human requests, but now embody some degree of autonomy and decision making. Some robots are even viewed to be “companions” to humans. As a result, a number of ethical issues have emerged. Thus this paper provides a survey of the current ethical landscape in robotics. Our goal is to show what has been done to date, with an eye toward the future. We believe that a lively and engaged discussion of ethical issues in robotics by roboticists and others, is essential for creating a better and more just world.

In this paper, we highlight the possible benefits, as well potential threats, related to the widespread use of robots. We follow the view that a robot cannot be analyzed on its own without taking into consideration the complex socio-technical nexus of today’s societies and that high-tech devices such as robots may influence how societies develop

in ways that could not be foreseen during the design of the robots. In our survey, we limit ourselves to presenting the ethical issues delineated by other authors and relay their lines of reasoning for raising the public’s concerns. We show that disagreements on what is ethical or not in robotics stems often from different beliefs on human nature and different expectations on what technology may achieve in the future. We stay away from offering a personal stance to these issues, so as to allow the reader to form her/his opinion.

In terms of robotic applications, we focus on service robots that peacefully interact with humans (Figures 1.(a) and 1.(b)) and on lethal robots created to fight on battlefields (Figures 1.(c) and 1.(d)). Other robotic applications are also discussed in the literature and have lead authors to emit a variety of concerns for our societies. Unfortunately, due to space constraints, we had to limit ourselves in our presentation. For instance, we omitted the question of unemployment caused by the development of industrial robots. This concern is in line with the general issue of using machines to replace human labor, a topic that is central to philosophical debates since the industrial revolution. Furthermore, we chose to not discuss the concerns that robots may one day be able to claim some social, cultural, ethical or legal rights, that robots may become sentient machines [56], which we would not longer be allowed to enslave [75], or the concern that we may create robots capable of annihilating mankind [17]. For a discussion on these issues, we refer the reader to [56], [75], [17].

II. WHO OR WHAT IS RESPONSIBLE WHEN ROBOTS CAUSE HARM?

Veruggio [100], [102] dates the beginnings of “roboethics” from two events. One was the *Fukuoka World Robot Declaration*, wherein it was stated that “next-generation robots will contribute to the realisation of a safe and peaceful society.” The other was the *Roboethics Roadmap* [101], which sought to promote a cross-cultural discussion among scientists to monitor the effects of the robotics technologies currently in use. More recently, an initial sketch of a *Code of Ethics* for the robotic community has been proposed [43]. This code offers general guidelines for ethical behavior. For example, the code reminds engineers that they may be held responsible for the actions taken by the artificial creatures that they helped to design. Along similar lines, Murphy and Woods [70] propose to rephrase the famous Asimovs Laws, which they view as robot-centric, in such a way as to remind robotics researchers and developers of their professional

Paweł Lichocki is with the Laboratory of Intelligent Systems, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland pawel.lichocki@epfl.ch

Peter H. Kahn, Jr. is with the Department of Psychology, University of Washington, Seattle, Washington 98195-1525 pkahn@u.washington.edu

Aude Billard is with the Learning Algorithms and Systems Laboratory, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland aude.billard@epfl.ch



(a)



(b)



(c)



(d)

Fig. 1. Robotic applications of service robots (a,b) and combat robots (c,d): (a) Childcare robot PaPeRo [32], [73]. [Photo courtesy of NEC Corporation. Unauthorized use not permitted.] (b) Paro therapeutic robot [89]. [Photo courtesy of AIST, Japan] (c) MQ-9 Reaper Hunter/Killer UAV by General Atomics Aeronautical Systems [33] (d) Modular Advanced Armed Robotic System (MAARS) by Foster-Miller [42].

responsibilities. For example, the first Law was replaced with: “A human may not deploy a robot without the human-robot work system meeting the highest legal and professional standards of safety and ethics” [70, p. 19].

All of the above implicates the responsibility ascription problem [69]: the problem of assigning responsibility to the manufacturer, designer, owner, or user of the robot or to the robot itself when use of a robot leads to a harmful event. From a philosophical perspective, it is generally agreed

that robots cannot themselves be held morally responsible [9], [25], [38] (although a few oppose this [95]), because computers as we conceive them today do not have intentionality [28]. From a psychological perspective, however, it remains an open question whether people include robots as an additional agent in the ascription of moral responsibility.

Who or what is responsible when robots cause harm (Figure 2)? Matthias [62] provides a seemingly simple answer. He argues that, in most cases, no one can be held accountable

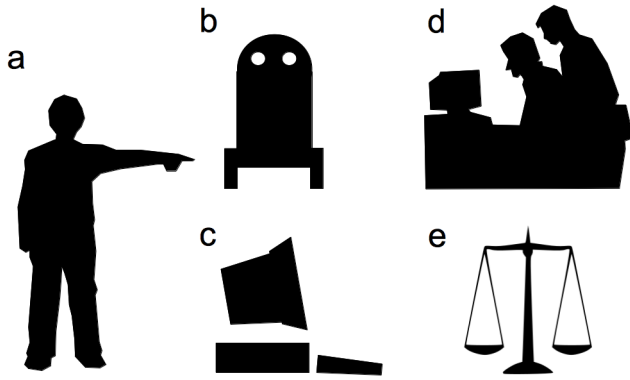


Fig. 2. (a) The *responsibility ascription* problem, i.e. the problem of assigning responsibility to either of the manufacturer, designer, owner or user of the machine when use of this machine led to an armful event is a yet largely open issue. (b) People may tend to blame the robots, because they falsely attribute to them moral agency [28]. (c) People blame the machine, even if they recognize the machine's lack of free will and lack of intentionality [30]. (d) Many ethicists argue that we should to some extent hold the engineers (the creators of the malfunctioning robots) responsible [60]. (e) In order to do so, we should use existing legal principles, or create new ones, if necessary [13].

for the robotic failures. For Matthias, the question of who is in control overcomes the responsibility ascription problem. Simply put, the greater the autonomy of the system, the less responsible the human. Further, Matthias reasons that with the advance of programming techniques (e.g. neural networks, evolutionary computation) that equip the agent with the ability to learn and hence to depart from its original program, it becomes impossible for the programmer to exhaustively test the behaviors of her creations. In other words, the programmer can no longer foresee all possible sets of actions that the robot may take when in function. Hence the programmer cannot be held responsible if harm should be done as a secondary effect of the robot interacting with humans, as long as the robot was not explicitly programmed to harm people. Matthias suggests that we should broadly adopt the idea of contracting insurances against harm caused by robots. Such new type of insurance would ensure that, when no-one can be held solely responsible for the harm done, then all the people involved in the incident would share the costs.

In contrast, Marino and Tamburini [60] state that Matthias's claims might have gone too far. In their opinion, determining who is controlling the robot cannot be a criterion (albeit even the unique criterion) to ascribe responsibility. They argue that engineers cannot be freed from all responsibility on the sole ground that they do not have a complete control over the causal chains implied by the actions of their robots [60]. They rather offer to use legal principles that are applied routinely for other purposes, so as to fill the "responsibility gap" that Matthias emphasized. They take the example of the legislation in place for ascribing responsibility to the "legally responsible" person when harm is done by the "dependent" person. As a result, parents can be held responsible for the act of their children, when they can

be found to have not provided adequate care or surveillance, even though there is no clear causal chain connecting them to the damaging events [60, pp. 49]. A similar solution is proposed by Asaro [13], who draws a parallel between robots and any other "completely unremarkable technological artifact[s]" (e.g. a toaster or a car). He shows that the Anglo-American civil law that rules for damages caused by these artifacts could also apply to damages produced by robots. For instance, if a manufacturer was aware of the danger that robots create, but failed to notify consumers, he may be charged with *failure to warn*. And even if the producer did not know about the danger, he could be accused of *failure to take proper care*, meaning that a manufacturer failed to recognize some easily foreseeable threat brought upon by his technology.

On the downside, Asaro points out that, while the civil law can relatively easily be extended to rule for robot use, the criminal law is hardly applicable to case of criminal actions caused by robots, as criminal actions can only be performed by *moral* agents. A moral agent is deemed so when it is recognized capable of understanding the moral concepts conveyed by the bylaws ruling our societies. Without moral agency, the act of wrongdoing is considered an accident and not a crime. Furthermore, only a moral agent can be punished and reformed. This assumes that the moral agent has the ability to develop and correct its concept of morality [13]. In this context, the responsibility ascription problem is hence reduced to the issue of attributing moral agency to the robot. Several authors have approached the problem of ascribing moral agency to robots. Solum [91]. For instance, Harnard proposes to use some sort of "moral" Turing tests to establish whether the robot can be held responsible in court [37].

Another issue around the responsibility ascription problem centers on attributing moral agency to a robot. In one study, Friedman and Millett [30] found that 83% of the undergraduate computer science majors they interviewed attributed aspects of agency - either decision-making and/or intentions - to computers. In addition, 21% of these students consistently held computers morally responsible for error. In another paper, Friedman and Kahn [28] identified a situation that may increase peoples attribution of agency to a machine: namely, when the machine is an expert recommendation system. Friedman and Kahn provide an example of the acute physiology and chronic health evaluation (APACHE) system [21]: a sophisticated computer-based modeling recommendation system to help hospital staff determine when to end life-support for patients in intensive care units. Friedman and Kahn argue that the more such a system is relied on for objective and authoritative information, the more difficult it becomes to override its recommendations, and the more likely staff, including physicians, could begin to attribute moral agency toward the system. As a potential solution to such problems, Friedman and Kahn offer two design strategies. First, computational systems should be designed in ways that do not denigrate the human user to machine-like status. Second, computational systems should be designed in ways that do not impersonate human agency

by attempting to mimic intentional states. The problem, however, in applying this second recommendation to robot design and implementation, especially those robots that have a humanoid form, is that such robots by design are conveying human attributes, and thus fostering this very problem.

III. ETHICAL ISSUES IN SERVICE ROBOTS

The design principle mentioned in the previous section aims at ensuring that robotic systems remain easily distinguishable from humans. Accordingly, this principle should help people ascribe responsibility in cases when the machine malfunctions or harms someone. However, as we noted, the current trend in robotics is the opposite, as there is a growing effort to design robots so that they look like humans [44], [45] or animals [31], [89].

The idea of designing machine masquerading humans was questioned by Miller on the ground of human freedom [67]. Miller argues that, if human-like robots really came to share human space on a daily basis, humans should be allowed to decide whether they wished or not to interact with these creatures; and if they should decide they wanted to interact solely with other humans, they should be given the freedom to do so. Similarly, efforts at endowing robots with social skills have been criticized on the ground that the number of meaningful social interactions that humans are typically capable to maintain is relatively small [23], [47]. Therefore, interacting with social artificial agents on a regular basis may lead people to become less prone to engage in social interactions with other people [66]. Others have suggested that people will likely form psychological intimacy with robots of the future, though of an impoverished form [50].

To shed some light on the above debate, researches have begun to investigate the type of human-robot relationships that arise when people interact with robotic systems that mimic human or animal behavior. In a series of four studies, for example, Kahn and his colleagues investigated children's social and moral relationships with the robot dog, the Artificial Intelligence roBOt (AIBO). The first three studies compared children's interaction with and reasoning about AIBO to, respectively, a stuffed (non-robotic) dog [49], a biologically live dog [65] and a mechanical non-robot dog [94], whereas the fourth study analyzed postings in AIBO online discussion forums that spoke of members' relationships with their AIBO [29]. Together, these four studies provide converging evidence that children and adults can and often do establish meaningful and robust social conceptualizations of and relationships with a robot that they recognize as a technology. For example, in the online discussion forum study, members affirmed that AIBO was a technology (75%), life-like (48%), had mental states (60%), and was a social being (59%).

Across these four studies, however, the researchers found inconsistent findings in terms of people's commitments to AIBO as a moral agent. In the online discussion forum study, for example, only 12% of the postings affirmed that AIBO had moral standing, including that AIBO had rights, merited respect, engendered moral regard, could be a recipient of

care, or could be held morally responsible or blameworthy [29]. In contrast, in the Melson et al. study [65] it was found that while on the one hand the children granted greater moral standing to a biologically live dog (86%) than to AIBO (76%), it was still striking that such a large percentage of children (76%) granted moral standing to the robot dog at all. One explanation for these inconsistent findings between studies is that the measures for establishing moral standing have been few, and themselves difficult to interpret. For example, two of the five moral questions in the Melson et al. study were: "If you decided you did not like AIBO anymore is it OK or not OK to throw AIBO in the garbage?" or "If you decided you did not like AIBO anymore is it OK or not OK to destroy AIBO?". The "Not OK" answers were interpreted as indicating moral standing. But plausibly one could make the same judgment about throwing away or destroying an expensive computer (because, e.g., it would be wasteful) without committing morally to the artifact [65].

Since humans can develop emotional attachment toward robots, concerns have been expressed regarding the long-term consequences that such attachment may have on the individual. This is especially relevant when the person is fragile, as it is the case with children and people with mental delays. However, there are also several reasons to believe that interacting with social robots may benefit some of these individuals [48], [54], [97]. For instance, interacting with robots that display social behavior may help children with autism acquire social skills [80], [26]. Robins et al. [80] conducted studies following children with autism interacting with a humanoid robot over the course of several weeks. Unknown to the children, the robot was puppeteered, so that it imitated the children's movement. Robins et al. showed that repeated exposure to the robot facilitated the emergence of spontaneous, proactive and playful behavior, which these children very rarely display. Furthermore, once accustomed to the robot, the children seemed to engage in more proactive interactive behavior with the adult investigator present in the room during the experiment. This leads, in some cases, to a triadic interaction between child, robot and adult. For example, children would acknowledge the presence of the investigator by spontaneously sitting on his lap for a few moments, holding his hand, or even trying to communicate by using simple words. However, it was not clear whether the social skills that children exhibited during the interactions with the robot had lasting effects.

In another study, Feil-Seifer and Matarić used a bubble-blowing robot in a triadic interaction of child-caretaker-robot. While the robot was not actually behaving socially, its automatic bubble-blowing behavior provoked more child-caretaker interactions. In a similar triadic child-parent-robot scenario, Kozima and colleagues conducted a series of studies using Keepon, a simple two-link robot-ball-face, whose motions conveyed emotional expressions. These studies supported Robins et al.'s findings that children with autism, in such triadic scenario, spontaneously displayed sociality and affect which they otherwise tend to avoid [55], [26]. In turn, Stanton and Kahn conducted a comparative study of children

with autism interacting with AIBO as opposed to a simpler mechanical toy dog [94]. Results showed that, in comparison to the toy dog, the children spoke more words to AIBO, and more often engaged in three types of behavior with AIBO typical of children without autism: verbal engagement, reciprocal interaction, and authentic interaction. In addition, there was highly suggestive evidence that the children, while in the AIBO session, engaged in fewer autistic behaviors. A survey of these studies can be found in [79].

As a whole, these studies seem to indicate that playing with robots that appear to behave in an autonomous and social manner may help children with autism display more of these social skills that autism therapy seeks to promote. Such a robotic aided therapy does not aim at developing attachment of the children towards the robot, but it might be a potential side-effect. The question remains if it is ethically correct to encourage children with autism to engage in affective interactions with machines incapable of emotions. Dautenhahn's and Werry's response is that, "from the perspective of a person with autism, and her needs, are these ethical concerns really relevant?" [24, pp. 35].

Similarly, robotic pets used in therapy with the elderly may offer some level of companionship. The seal robot, Paro, is probably the best example of such an application [89] (Figure 1.(b)). Wada et al. [104] report on extended use of Paro as part of therapeutic sessions in pediatric wards and elderly institutions world-wide. Results showed that interaction with Paro improved the patients' and elderly people's moods and reduced their stress level [103]. It made them more active and communicative both among themselves and with their caretakers. A pilot study using electroencephalography (EEG) suggested that this robot therapy may improve pattern of brain activity in patients suffering from dementia [104]. Furthermore, the effects of long-term interaction between Paro and the elderly was found to last for more than a year [105].

While the above results speak in favor of using robots for therapy with the elderly, Sharkey offers a more cautious argumentation [85]. In his opinion such surrogate companions do not really alleviate the elderly's isolation and people are deluded about the real nature of their relationship to the devices [92] (Figure 3). Furthermore, even the robots that are clearly helping the elderly to maintain independence in their own homes [27] (e.g. robots used to remind the patient to take her medication), could lead to a situation where the elderly is left exclusively to the care of machines, and deprived of the benefits of human contact, which is often provided by caregivers [93]).

Robot-nannies are another example of robotics applications that raise ethical questions [88]. There is an effort, mainly in South Korea and Japan, to build more sophisticated robots that could not only monitor babies (as e.g. Personal Partner Robot by NEC [32], Figure 1.(a)) but that would also be equipped with enough autonomy so as to call upon human caretakers only in unusual circumstances. It is likely that children will spend time playing with childcare robots, as researchers progress in designing ways for the robot to offer



Fig. 3. Interacting with robots that display social behavior may help children with autism-acquired social skills. The question remains whether it is ethically correct to encourage children with autism to engage in affective interactions with machines incapable of emotions. But, from the perspective of a person with autism, and her needs, are these ethical concerns really relevant? [24, pp. 35]. In a broader context, some believe that, the surrogate companions (e.g. robots assisting the elderly) are becoming more common, because people are deluded about the real nature of their relationship to the devices [95]. (Photo courtesy of KASPAR robot by University of Hertfordshire [108])

sustained interactions with the child that may span months or even years [51], [63], [88]. The results, however, may be detrimental to the physical and mental development of the child if children were to be left without human contact for many hours per day [85]. This remains very speculative as the psychological impact that such robotics care may have on children' development is unknown. Some of the literature has attempted to draw parallels with reports on severe social disfunctions in young monkeys who interacted solely with artificial caretakers throughout the first years of development [61], [16], [88]. Perhaps of more pressing concern is the fact that there is no regulation to specifically deal with the case of child abuse when the child is cared for by a robot (national and international laws protecting children from mistreatment such as The United Nations Convention on the Rights of Child [71] do not cover this case) [88]. While one may argue that, when the time will really come to see robots caring for children, one will work on the associated legal issues, some people counter that this may be a bigger challenge than expected, as providing a unified code of ethics for regulating the use of robot-nannies may be impossible due to cultural differences between nations [36].

IV. ETHICAL ISSUES IN LETHAL ROBOTS

In Section III, we covered some of the ethical issues that stem from current or foreseen robotic applications of service robots for education and therapy. Of equal if not more immediate ethical concern are the current military applications of robots. Even though fully autonomous robots are not yet running in battlefields, the risks and benefits that introducing such autonomous lethal machine may have on wars are of crucial importance. Furthermore, because military technology often finds its way into civilian applications, such as security or policing [14], [87], ethical issues related to military robots impacts society on even broader level.

Currently, the decision to use a robotic device to kill human beings still lies with a human operator. This decision stems not out of any technical necessity, but from the desire to make sure that the human remains “in the loop” [14]. It is clear that the margin that separates us from having fully autonomous armed systems in the battlefield is thinning. Even if all armed robots were to be supervised by humans, one may still wonder to what extent the human is still in control [9]. Moreover, there may be cases where one cannot avoid giving full autonomy to the system. For instance, combat aircrafts must be fully autonomous in order to effectively operate [99]. Sharkey predicts that, as the number of robots in operation in the battlefield increases, the robots may outnumber human soldiers. He then argues that it will become impossible for humans to simultaneously operate all these robots. Robots will then have to be fully autonomous [83].

One ethical issue (perhaps the issue that received most attention to date) arising from increasing autonomy of war robots has to do with the problem of discriminating between enemy combatants and innocent people. This distinction is at the core of the “just war theory” [106] and the humanitarian laws [82]. These laws stipulate that only enemy combatants are legitimate targets and prohibit attacks against any other non-legitimate targets [84], [14]. Sharkey argues rightfully that our robots are still far from having visual capabilities that may allow to discriminate faithfully between legitimate and non-legitimate targets, even in close-contact encounter [85] (Figure 4). In addition, distinguishing between legitimate and illegitimate targets is not purely a technical issue and is further complicated by the lack of a clear definition of what counts as a civilian¹. But even if one was provided with a precise definition that could be encoded in a computer program, it is doubtful that robots would achieve, in the foreseeable future, a level of complexity in robot cognition that would allow the robot to recognize ambiguous situations involving a non-legitimate target manipulating lethal instruments (such as, for example, a situation where a child is carrying guns or ammunition). Sharkey argues that autonomous lethal systems should not be used until one can fully demonstrate that the systems can faithfully distinguish between a soldier and a civilian in all situations [83]. Others, however, argue that this

condition is too stringent, since even humans make errors of this kind [58] (Figure 4). Arkin, counters that, although unmanned robotic systems may make mistakes, it would on average behave more ethically than human beings [9]. In support, Arkin cites a report from the Surgeon General’s Office [96] regarding the ethics of soldiers. Less than half of the soldiers believed that non-fighters should be treated with dignity. The other half was unclear as to how they should be treated. Moreover, one tenth of the soldiers had mistreated civilians and one third reported having at least once faced a situation where they felt incapable of deciding from an ethical stance the correct action (although all had received ethical training). Arkin argues that since human soldiers appear to misbehave from time to time, using machines that are more reliable and hence would, on average, make less mistakes, should bring more good than harm. Lin and colleagues share the view that human soldiers are indeed less reliable and report on evidence that human soldiers may act irrationally when in fear or stressed. They hence concur that combat robots, that are affected neither by fear or stress, may act more ethically than human soldiers irrespective of the circumstances [58].

Lin and colleagues point to one more issue related to using combat robots. As in the case of any other new computational technology, errors and bugs will inevitably exist and these will lead combat robots to cause harmful accidents [58]. Such bugs or errors will be far more costly as human lives might be at stake. They advise extensive testing of each military robot prior to usage. Nevertheless, they anticipate that, due to their high complexity, these robots will still occasionally make errors and kill innocent people [58]. Such errors could even lead to accidental wars, if the robot’s unexpected aggressive behavior was to be interpreted by the opponent as an act of war [14]. Groups of people interested in starting a war may seize upon such accidents to justify hostilities.

Even if one is not disputing the ethical question of entering in war, one may want to question the ethics of having armed robots fully autonomous and used routinely in battlefields, especially when only one side may have robots. Politicians may favor efforts to replace human fighters with robots, as each country feels a moral obligation to protect the lives of its soldiers [83]. However, there may be long-term consequences of waging these so-called risk-free wars² or push-button wars³. Since such wars will return wrecked metal in place of dead bodies (at least to the country using only robots), the emotional impact that wars currently have on civilians of that country will be largely lessened. The above is true only for the civilians not affected directly by combat, i.e. for wars fought in a distance.

It is feared that this may make it easier for a country to launch a war. These wars may also last for longer periods of time [58]. There are contradicting opinions whether this may

¹The 1944 Geneva Convention advises to use common sense and the 1977 Protocol 1 defines a civilian as any person who is not a fighter. [72]

²A war where pilotless aircraft can beat a country’s forces before sending in the ground robots to clean up” [83, p. 16]

³A war in which the enemy is killed at a distance, without any immediate risk to oneself [14, pp. 62]

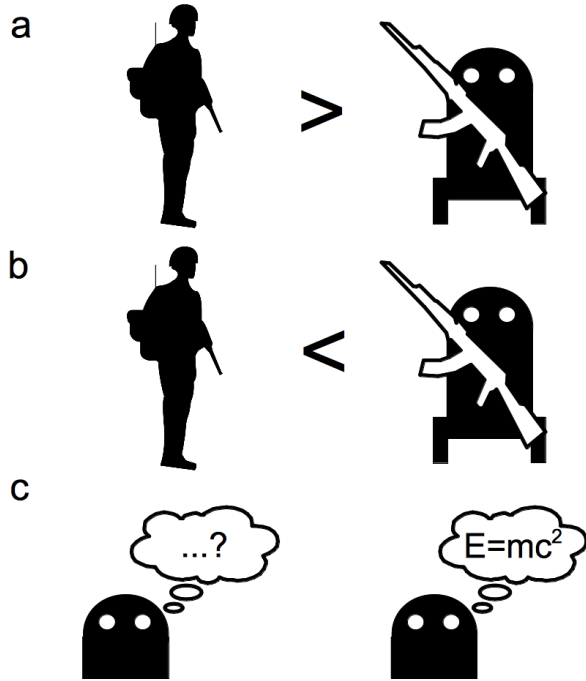


Fig. 4. (a) Noel Sharkey argues that the cognitive capabilities of robots do not match that of humans, and thus lethal robots are unethical, as they may make mistakes more easily than humans [85]. (b) Ronald Arkin believes that although an unmanned system will not be able to perfectly behave in battlefield, it can perform more ethically than human beings [9]. (c) In part, the question about the morality of using robots in the battlefield involves commitments on the capability of artificial intelligence. (Photo of the soldier's silhouette by Ruminglass and Quibik, under CC BY-SA 3.0)

result in people growing indifferent to the conduct of war. Sharkey fears that this would be the case [83], whereas Asaro believes that people are nearly always averse to starting an unjust war, irrespective of whether or not it would lead to human fatalities [14, pp. 58]. The fact that the war is risk-free does not by itself make it more acceptable [14]. Lin and colleagues counterweight this line of reasoning, arguing that such reasoning may lead to even more “dangerously foolish” ideas, such as the idea of trying to prevent wars to happen by increasing the brutality of the fighting [58].

It was also argued that risk-free wars might increase terrorism, as the only possibility to strike back at a country who uses mainly robots in wars is to attack its citizens [83]. The less advanced, “technologically speaking”, side may advocate terrorism as a morally acceptable means to counterattack, on the ground that “robot armies” are the product of a rich and elaborate economy, and that members of that economy are the next-best legitimate targets [14, pp. 64]. Hence, risk-free wars may paradoxically increase the risk for civilians [46]. However, Asaro reminds us that wars are deemed morally acceptable as long as they do not harm civilians [14]. According to this definition, terrorism would not be justified, irrespective of whether it is meant as a response to a country using robot armies. Thus, the fear that terrorism may increase as a result of using robot armies does not constitute, in Asaro’s view, a valid moral objection

to using robot armies. Only the questions of whether robot armies cause more harm or a greater injustice (than using a human army) are of essence in the debate [14].

In contrast, Arkin anticipates that we will not end-up with armies of unmanned systems operating on their own, but that rather heterogeneous teams composed of autonomous systems and human soldiers will work together on the battlefield. Wars would hence not be fully risk-free and so the dreaded consequences in increased terrorism or in societal indifference are not to be feared. Furthermore, Arkin expects that mixed teams, composed of robots and human soldiers, will act more ethically than any group composed of solely human soldiers. One reason is that robots equipped with video cameras (or other sensors) will record and report actions on the battlefield. Thus, they might serve as a deterrent against unethical behavior, as such acts would be registered. However, Lin and colleagues argue that if soldiers were to know that they are being watched by their fellow robot soldiers, they may no longer trust them and this could impact team cohesion and armed effectiveness [58].

In terms of law, Sharkey points out that the legal status of war robots is unclear [86]. For example, while the unmanned aerial vehicle RQ-1 Predator (Figure 1.(d)) was developed as a reconnaissance machine (hence the R in the name), it was subsequently equipped with Hellfire missiles and renamed MQ-1 (where M stands for multipurpose). The MQ-1 was never approved as a weapon; however, it did not need to be. The reason is as follows. Because the bare RQ-1 was not considered a weapon (since it was meant for surveillance only) and the hellfire missiles have already been approved separately as weapons, the combination did not need special approval [19]. Such reasoning may create a precedent whereby armed robots with growing level of autonomy can be created and used with little legal control. Asaro notices that “what is and what is not acceptable in war” is ultimately the subject of *convention* between nations [14, pp. 64]. He argues that we can find support in existing laws only to a certain extent. Eventually, the international community will be forced to create new laws and treaties in order to regulate the use of autonomous fighting robots.

V. MACHINE ETHICS

Although still in its early stages, machine ethics offers a practical approach to introduce ethics in the design of autonomous machines. Machine ethics aims at giving the machine some autonomy, while ensuring that its behavior will abide ethical rules. Primarily, machine ethics seeks methods to ensure that the machines’s behavior toward humans is proper [4], but it also may extend to designing rules driving ethical behavior of a machine toward another machine [6]. Machine ethics extends the field of computer ethics that is concerned with how people behave with their computers to address the problem of how machine behave in general [2].

The interest in machine ethics is driven by the fact that robots have been already tightly integrated into human societies. Thus, since the robots already interact with humans and, as argued in Section II, engineers could be

held responsible (to certain extent) for the actions of their creations, it is desirable to find methods of equipping the machines with moral behavior. Importantly, although the public attention might be focusing on the military application (such as Arkin's military advisor providing guidance on the use of lethal force by a robot [11]), machine ethics seems to be more concerned with service robots. The examples of such applications are many. Robots that share the workbench with humans in the industry might no longer be considered just a manufacturing tool, but also a "colleague" with whom workers interact [20]. Artificial sales-agents in e-commerce, which can predict customers behaviors, should not abuse this knowledge by displaying unethical behavior [39]. Driverless trains in extreme situations might be forced to make decisions that could have life or death implications [2].

Asimov's Laws of Robotics are one of the first and best known proposal to embed ethical concepts in the controller of the robot⁴. According to these, all robots should under all circumstances obey three laws:

- 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- 2) A robot must obey orders it receives from human beings, except when such orders conflict with the first law.
- 3) A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

Later, Asimov added the 4th Law (known as the Law Zero)

- 4) No robot may harm humanity or, through inaction, allow humanity to come to harm.

Many researchers recognize that the Asimov's laws assumes that robots have sufficient cognition to make moral decisions in all situations, including the complicated ones, in which even humans might have doubts [70]. Consequently, keeping in mind the current level of AI, these laws, although simple and elegant, serve no useful practical purpose [9] and are thus viewed as an unsatisfactory basis for machine ethics [8], [34]. Nevertheless, Asimov's laws often serve as a reference or starting point in the discussions related to machine ethics.

Fedaghi [1] proposes a classification scheme of ethical categories to simplify the process by which a robot may determine which action is most ethical in complicated situations. As a proof of concept, Fedaghi applies this classification to decompose Asimov's laws, showing hereby that these laws, once rephrased, can support logical reasoning. Such an approach is in-line with so-called *procedural ethics* [59], which develops procedures to guide the process by which ethical decisions are made [1]. A similar approach is presented in [18] that draws inspiration in Leibniz's dream of a universal moral calculus [57]. There, deontic logic [22], [68] (i.e. logic extended with special operators for representing ethical concepts) is used instead of Asimov's laws to ground the robot's

ethical reasoning. Such a methodology aims at maximizing the likelihood that a robot will behave in a certifiably ethical fashion. That is, the robot's actions will be determined so that the ethical correctness of the resulting robot's behavior can be ensured through formal proofs. Such formal proofs check if a given robot (a) only takes permissible actions and (b) performs all obligatory actions (subject to ties and conflicts) [12]. Promoters of such methodology reason that human relationships and, by extension human-robot relationships, need to be based on some level of trust [107]. Such a formal and logical approach to describing robot behavior may help to determine whether the system is trustworthy or not. In contrast, they view inductive reasoning, that is based on case studies, as unreliable, because, while the "premise (success on trials) may all be true, the conclusion (desired behavior in the future) might still be false" [18], [90].

Others oppose this point of view and advocate the use of case-based reasoning (CBR) [74]. They reason that people can behave ethically without learning ethics (drawing a parallel to the fact that one can speak fluently a language without having received any formal grammar lessons) [81]. For example, McLaren implemented a CBR ethical reasoner [64] and Anderson created a machine learning system that automatically derives rules (principles) from cases provided by an expert ethicist [3], [7], [5]. For example, Arkin uses deliberative/reactive autonomous robotic architectures and provides the theory and formalisms for ethical control [10] and applies these to automatic military advisor [11]. He considers stimuli to behavior mappings and extends them with ethical constraints in order to ensure appropriate robot response (consistent with law). In the another example, Honarvar [40] used a CBR-like mechanism to train an artificial neural network (ANN) to classify what is morally acceptable in a believe-desire-intention (BDI) framework [77]. He used this framework to augment the ethical knowledge of sales-agent in an e-commerce application [39].

A particular machine ethics system that is comparatively easy to implement is one based on utilitarianism. It uses a mathematical calculus to determine the best choice (by computing and maximizing the "goodness", however defined, of all actions) [4]. But, while utilitarianism values benefits brought upon society as a whole, it does not protect the fundamental rights of each individual [78], [11] and thus is of limited interest [35]. Still, practical work with a certain utilitarian flavor can be found in the literature, as most CBR systems presented previously assume that an arithmetic value is the main basis for determining what it is moral to do [53].

The last approach we will mention is the rule-based one proposed by Powers [76]. Powers argues that ethical systems such as Kant's categorical imperative⁵ lead naturally to a set of rules. This approach hence assumes that an ideological ethical code can be translated into a set of core rules. This is somewhat similar to the deontic logic we reviewed

⁴The Asimov's Law of Robotics were first introduced in the short science-fiction story Runaround [15]

⁵A categorical imperative denotes an absolute, unconditional requirement that asserts its authority in all circumstances, e.g. "Act only according to that maxim whereby you can at the same time will that it should become a universal law" [52, pp.30].

earlier on. It allows robots to logically derive new ethical rules, appropriate to particular and new situations. Although interesting, this approach has not gathered much attention, as researchers usually turn to pure logic systems or case-based reasoning. Also, Powers' ethical system had been criticized by Tonkens [98] on the basis that the development of Kantian artificial agents is itself against Kant's ethics. According to Kant, moral agents are both rational and free, whereas machines can only be rational. Hence, the mere fact of implementing a sense of morality into machines limits the machine's freedom of thought and reasoning.

In brief, machine ethics is composed of a number of interesting attempts to embed ethical rules in the robot's controller. These rules may be common or popular in society, such as Asimov's laws, or they may be derived from classical philosophical approaches, such as utilitarianism or Kantian ethics. Logical reasoning is the driving framework for most approaches. While still in infancy, machine ethics is a valuable attempt to provide robots with ethical behavior. But the approach may fall prey to several problems discussed throughout this paper. Three of them stand out. One, if machines are not capable of being moral agents, as most philosophers agree, then it seems suspect to design into them the ability to make moral decisions. Second, equipping the machines with morality (assuming it is possible), does not need to be a moral act on its own and might depend on the application one has in mind while developing a moral robot. For example, embedding morality into robonnies or combat robots could lead to the widespread use of them, which could have severe negative consequences on the society (see Section III and Section IV). Finally, in an attempt to embed ethics into machines, due to their limited cognition, one must often unduly simplify the moral life. This seems to stand against the very goal of machine ethics itself (at least to some extent). It seems it is still too early to judge if the methods of machine ethics will prove useful or not, and awaits more applications implemented in life.

VI. CONCLUSIONS

Almost everyone agrees that they want robots to contribute to a better world, and a more ethical one. The disagreements arise in how to bring that about. Some people want to embed ethical rules in the robots controller, and employ such robots in morally challenging contexts, as on the battlefield. Others argue vehemently against this approach: that robots themselves are incapable of being moral agents and thus should not be designed to have moral-decision making abilities. Others want to leverage the social aspects of robotics in bringing about human good. Along these lines, researchers have explored how robots can help children with autism, or assist the elderly physically and thereby provide the elderly with enough autonomy to allow them to live in their own residences. Other researchers have explored how robots can provide companionship for the elderly, and for the general population. Still others have worried that no matter how sophisticated robots become in their form and function, their technological platform will always separate people from

them, and prevent depth and authenticity of relation from forming.

These are all open questions. Some are philosophical in nature, as is the question of whether robots are moral agents, or could be in the future. Some are psychological, as in the question of whether people attribute moral responsibility to robots that cause harm. Some require political answers and new legislation. And, finally, some - if not many - of the questions require thoughtful and on-going responses by those of us who are the engineers and designers of the robots. The engineer is no longer entirely free of responsibility regarding the ethical consequences of his/her creation. This seems at odds with the way research is currently done in robotics. Rarely does one question the long-term ethical consequences of the research reported upon in scientific publications⁶. There are several reasons for this. On the one hand, most of these damaging long-term consequences seem very speculative and still far away from the technological reality. On the other hand, it is expected that these issues will be disputed at a political level and hence that it is perhaps not the role of the engineers and scientists to discuss these.

Some scientists however discuss these issues, but, as with any debate, people have sometimes opposite views on which robotic application is ethical and which is not. We showed that such dissensions stemmed often from different beliefs on human nature and different expectations on what technology may achieve in the future. While it is difficult to anticipate how and when robots will come to play an active role in our society, there is no reason why one should not continue discussing various scenarios. We might be motivated by the beauty of our artifacts. Or by their usefulness. Or by the economic rewards. But in addition we are morally accountable for what we design and put out into the world.

VII. ACKNOWLEDGMENTS

This work was partially supported by (a) the Swiss National Science Foundation (grant number K-23K0-117914) and (b) the European Commission under contract number FP7-248258 (First-MM) and (c) the National Science Foundation in the United States (grant number IIS-0905289).

REFERENCES

- [1] S.S. Al-Fedaghi. Typification-based ethics for artificial agents. In *2nd IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pages 482–491, 2008.
- [2] C. Allen, W. Wallach, and I. Smit. Why Machine Ethics? *IEEE Intelligent Systems*, 21(4):12–17, 2006.
- [3] M. Anderson, S.L. Anaderson, and C. Armen. MedEthEx: A prototype medical ethics advisor. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 1759–1765, 2006.
- [4] M. Anderson, S. Anderson, and C. Armen. Towards machine ethics: Implementing two action-based ethical theories. In *Technical Report FS-05-06 of AAAI Fall Symposium on Machine Ethics*, pages 1–7, 2005.

⁶We are not referring here to short-term ethical consequences of a research, such as a research that involves human subjects. Clearly, these are always carefully scrutinized and this research must be approved by ethical committee prior to the conduct of the project.

- [5] M. Anderson and S. L. Anderson. Ethical Healthcare Agents. In *Advanced Computational Intelligence Paradigms in Healthcare - 3* (Stud. Comput. Intell., 107:233–257), M. Sordo, S. Vaidya, and L. C. Jain, Eds. Berlin: Springer-Verlag, 2008.
- [6] M. Anderson and S.L. Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–27, 2007.
- [7] M. Anderson, S.L. Anderson, and C. Armen. An approach to computing ethics. *IEEE Intelligent Systems*, 21(4):56–63, 2006.
- [8] S.L. Anderson. Asimov’s “three laws of robotics” and machine metaethics. *AI & Society*, 22(4):477–493, 2008.
- [9] R.C. Arkin. Governing ethical behavior: embedding an ethical controller in a hybrid deliberative-reactive robot architecture - Part I: motivation and philosophy. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 121–128, 2008.
- [10] R.C. Arkin. Governing ethical behavior: embedding an ethical controller in a hybrid deliberative-reactive robot architecture - Part II: formalization for ethical control. In *Proceedings of the 1st Conference on Artificial General Intelligence*, pages 51–62, 2008.
- [11] R.C. Arkin. Governing lethal behavior: embedding ethics in a hybrid deliberative-reactive robot architecture - Part III: representational and architectural considerations. In *Technology in Wartime Conference*, 2008. Stanford Law School [Online]. Available: <http://hdl.handle.net/1853/22715>
- [12] K. Arkoudas, S. Bringsjord, and P. Bello. Toward ethical robots via mechanized deontic logic. In *Technical Report FS-05-06 of AAAI Fall Symposium on Machine Ethics*, pages 17–23, 2005.
- [13] P. Asaro. Robots and responsibility from a legal perspective. Presented at *Workshop on Roboethics, IEEE International Conference on Robotics and Automation (ICRA)*, 2007 [Online]. Available: http://www.roboethics.org/icra2007/contributions/ASARO_LegalPerspectives.pdf
- [14] P. Asaro. *How just could a robot war be?*, pp. 50–64. Amsterdam: IOS Press, 2008.
- [15] I. Asimov. Runaround. *Astounding Science Fiction*, 1942.
- [16] D. Blum. *Love at Goon Park: Harry Harlow and the science of affection*. Basic Books, New York, 2002.
- [17] N. Bostrom. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, [Online], 9(1), 2002. Available: <http://www.jetpress.org/volume9/risks.html>
- [18] S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.
- [19] J. Canning, G. Riggs, O. Holland, and C. Blakelock. A concept for the operation of armed autonomous systems on the battlefield. In *Association for Unmanned Vehicle Systems International Annual Symposium and Exhibition*, Anaheim, CA, 2004.
- [20] B. Çürüklü, G. Dodig-Crnkovic, and B. Akan. Towards industrial robots with human-like moral responsibilities. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 85–86. ACM, 2010.
- [21] R.W.S. Chang, B. Lee, S. Jacobs, and B. Lee. Accuracy of decisions to withdraw therapy in critically ill patients: clinical judgment versus a computer model. *Critical Care Medicine*, 17(11):1091–1097, 1989.
- [22] B.F. Chellas. *Modal logic: an introduction*. Cambridge University Press, Cambridge, 1980.
- [23] K. Dautenhahn. Robots we like to live with?!-a developmental perspective on a personalized, life-long robot companion. In *13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, pages 17–22, 2004.
- [24] K. Dautenhahn and I. Werry. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, 12(1):1–35, 2004.
- [25] D. Dennett. *When HAL kills, who’s to blame?*, chapter 16. MIT Press, 1996.
- [26] D. Feil-Seifer and M. Matarić. Robot-assisted therapy for children with autism spectrum disorders. In *Proceedings of the 7th International Conference on Interaction Design and Children*, pages 49–52, 2008.
- [27] J. Forlizzi, C. DiSalvo, and F. Gemperle. Assistive robotics and an ecology of elders living independently in their homes. *Human-Computer Interaction*, 19(1):25–59, 2004.
- [28] B. Friedman and P.H. Kahn Jr. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1):7–14, 1992.
- [29] B. Friedman, P.H. Kahn Jr, and J. Hagman. Hardware companions?: What online AIBO discussion forums reveal about the human-robotic relationship. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 273–290, 2003.
- [30] B. Friedman and L. Millet. “It’s the computer’s fault”: reasoning about computers as moral agents. In *Conference Companion on Human Factors in Computing Systems*, pages 226–227, 1995.
- [31] M. Fujita. AIBO: Toward the era of digital creatures. *The International Journal of Robotics Research*, 20(10):781–794, 2001.
- [32] Y. Fujita, S.I. Onaka, Y. Takano, J.U.N.I. Funada, T. Iwasawa, T. Nishizawa, T. Sato, and J.U.N.I. Osada. Development of childcare robot PaPeRo. *Nippon Robotto Gakkai Gakujutsu Koenkai Yokoshu (CD-ROM)*, 23:1–11, 2005.
- [33] General Atomics Aeronautical. Predator B, [Online]. Available: http://www.ga-asi.com/products/aircraft/predator_b.php
- [34] J. Gips. *Towards the ethical robot*, pages 243–252. MIT Press, 1995.
- [35] C. Grau. There is no “I” in “robot”: robots and utilitarianism. *IEEE Intelligent Systems*, 21(4):52–55, 2006.
- [36] S. Guo and G. Zhang. Robot rights. *Science*, 323(5916):876, 2009.
- [37] S. Harnad. Minds, machines and Turing. *Journal of Logic, Language and Information*, 9(4):425–445, 2000.
- [38] K.E. Himma. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1):19–29, 2009.
- [39] A.R. Honarvar and N. Ghasem-Aghaee. Towards an ethical sales-agent in e-commerce. In *2010 International Conference on e-Education, e-Business, e-Management and e-Learning*, pages 230–233, 2010.
- [40] Honarvar, A.R. and Ghasem-Aghaee, N. An artificial neural network approach for creating an ethical artificial agent. In *Proceedings of the 8th IEEE International Conference on Computational Intelligence in Robotics and Automation*, pages 290–295, 2009.
- [41] IFR Statistical Department. Executive Summary of World Robotics 2009 Industrial Robots and Service Robots, [Online]. Available at: http://www.worldrobotics.org/downloads/2009_executive_summary.pdf
- [42] Foster-Miller Inc. TALON Family of Military, Tactical, EOD, MAARS, CBRNE, Hazmat, SWAT and Dragon Runner Robots, retrieved at 17.09.2010.
- [43] B. Ingram, D. Jones, A. Lewis, M. Richards, C. Rich, and L. Schachterle. A code of ethics for robotics engineers. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, pages 103–104, 2010.
- [44] H. Ishiguro. Android science: Conscious and subconscious recognition. *Connection Science*, 18(4):319–332, 2006.
- [45] H. Ishiguro. Interactive humanoids and androids as ideal interfaces for humans. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, pages 2–9, 2006.
- [46] P.H. Kahn Jr. The Paradox of Riskless Warfare. *Philosophy & Public Policy Quarterly*, 22(3):2–8, 2002.
- [47] P.H. Kahn Jr, N.G. Freier, B. Friedman, R.L. Severson, and E. Feldman. Social and moral relationships with robotic others. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, pages 545–550, 2004.
- [48] P.H. Kahn Jr, B. Friedman, and J. Hagman. I care about him as a pal: Conceptions of robotic pets in online aibo discussion forums. In *CHI’02 extended abstracts on Human Factors in Computing Systems*, pages 632–633, 2002.
- [49] P.H. Kahn Jr, B. Friedman, D.R. Perez-Granados, and N.G. Freier. Robotic pets in the lives of preschool children. In *CHI’04 Extended Abstracts on Human Factors in Computing Systems*, pages 1449–1452, 2004.
- [50] P.H. Kahn Jr, J.H. Ruckert, T. Kanda, H. Ishiguro, A. Reichert, H. Gary, and S. Shen. Psychological intimacy with robots?: using interaction patterns to uncover depth of relation. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pages 123–124, 2010.
- [51] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1 & 2):61–84, 2004.
- [52] I. Kant (translated by J.W. Ellington). *Grounding for the Metaphysics of Morals 3rd ed.* Hackett Pub. Co., 1993 (written in 1785).
- [53] M. Keefer. *Moral reasoning and case-based approaches to ethical instruction in science*, volume 19 of *Science & Technology Education Library*, pages 241–259. Springer, 2003.
- [54] C. Kidd, W. Taggart, and S. Turkle. A sociable robot to encourage social interaction among the elderly. In *Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1050–1059, 2006.

- [55] H. Kozima, C. Nakagawa, and Y. Yasuda. Children-robot interaction: a pilot study in autism therapy. In *From Action to Cognition* (Progress in Brain Research, 164:385–400), C. von Hofsten and K. Rosander, Eds., 2007.
- [56] R. Kurzweil. *The singularity is near: When humans transcend biology*. Viking Penguin: New York, 2005.
- [57] G.W. Leibniz (translated by G.M. Ross). *Notes on Analysis*. Past Masters. Oxford Univ. Press, 1984.
- [58] Lin, P. and Bekey, G. A. and Abney, K. Robots in war: Issues of risk and ethics. in *Ethics and Robotics*, Amsterdam: IOS Press, pages 49–67, 2009.
- [59] W. Maner. Heuristic methods for computer ethics. *Metaphilosophy*, 33(3):339–365, 2002.
- [60] D. Marino and G. Tamburrini. Learning robots and human responsibility. *International Review of Information Ethics*, 6:46–50, 2006.
- [61] W.A. Mason and G. Berkson. Effects of maternal mobility on the development of rocking and other behaviors in rhesus monkeys: A study with artificial mothers. *Developmental Psychobiology*, 8(3):197–211, 1975.
- [62] A. Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004.
- [63] N. Mavridis, C. Datta, S. Emami, A. Tanoto, C. BenAbdelkader, and T. Rabie. FaceBots: robots utilizing and publishing social information in facebook. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 273–274, 2009.
- [64] B.M. McLaren. Computational models of ethical reasoning: challenges, initial steps, and future directions. *IEEE Intelligent Systems*, 21(4):29–37, 2006.
- [65] G.F. Melson, P.H. Kahn Jr, A. Beck, and B. Friedman. Robotic pets in human lives: implications for the human-animal bond and for human relationships with personified technologies. *Journal of Social Issues*, 65(3):545–567, 2009.
- [66] G.F. Melson, P.H. Kahn Jr, A.M. Beck, B. Friedman, T. Roberts, and E. Garrett. Robots as dogs?: children’s interactions with the robotic dog AIBO and a live australian shepherd. In *CHI’05 extended abstracts on Human Factors in Computing Systems*, pages 1649–1652, 2005.
- [67] K.W. Miller. It’s not nice to fool humans. *IT Professional*, 12(1):51–52, 2010.
- [68] J.H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.
- [69] P.A. Mudry, S. Degallier, and A. Billard. On the influence of symbols and myths in the responsibility ascription problem in roboethics - a roboticist’s perspective. In *The 17th IEEE International Symposium on Human Robot Interactive Communication (RO-MAN)*, pages 563–568, 2008.
- [70] R. Murphy and D.D. Woods. Beyond Asimov: the three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4):14–20, 2009.
- [71] United Nations. Convention on the Rights of the Child. *Treaty Series*, 1577:3, 1989.
- [72] International Committee of the Red Cross. Protocol additional to the geneva conventions of 12 august 1949 (article 50), 1977.
- [73] J. Osada, S. Ohnaka, and M. Sato. The scenario and design process of childcare robot, PaPeRo. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, page 80, 2006.
- [74] S.K. Pal and S. Shiu. *Foundations of soft case-based reasoning*. Wiley-Interscience, 2004.
- [75] S. Petersen. The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):43–54, 2007.
- [76] T.M. Powers. Prospects for a Kantian machine. *IEEE Intelligent Systems*, 21(4):46–51, 2006.
- [77] A.S. Rao and M.P. Georgeff. BDI agents: from theory to practice. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS)*, pages 312–319, 1995.
- [78] J. Rawls. *A theory of justice*. Harvard University Press, Cambridge, 1999.
- [79] D.J. Ricks and M.B. Colton. Trends and considerations in robot-assisted autism therapy. In *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4354–4359, 2010.
- [80] B. Robins, K. Dautenhahn, R.T. Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120, 2005.
- [81] R. Rzepka and K. Araki. What could statistics do for ethics? The idea of a commonsense processing based safety valve. In *Technical Report FS-05-06 of AAAI Fall Symposium on Machine Ethics*, pages 85–87, 2005.
- [82] M.N. Schmitt. The Principle of discrimination in 21st century warfare. *Yale Human Rights and Developmental Law Journal*, 2(1):143–164, 1999.
- [83] N. Sharkey. Cassandra or false prophet of doom: AI robots and war. *IEEE Intelligent Systems*, 23(4):14–17, 2008.
- [84] N. Sharkey. Grounds for discrimination: autonomous robot weapons. *RUSI Defence Systems*, 11(2):86–89, 2008.
- [85] N. Sharkey. The ethical frontiers of robotics. *Science*, 322(5909):1800–1801, 2008.
- [86] N. Sharkey. Death strikes from the sky: the calculus of proportionality. *IEEE Technology and society magazine*, 28(1):16–19, 2009.
- [87] N. Sharkey. The robot arm of the law grows longer. *Computer*, 42(8):116,113–115, 2009.
- [88] N. Sharkey and A. Sharkey. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–190, 2010.
- [89] T. Shibata, M. Yoshida, and J. Yamato. Artificial emotional creature for human-machine interaction. In *1997 IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 2269–2274, 1997.
- [90] B. Skyrms. *Choice and chance: An introduction to inductive logic*. Wadsworth Publishing, 1999.
- [91] L.B. Solum. Legal personhood for artificial intelligences. *North Carolina Law Review*, 70(1):1231–1287, 1992.
- [92] R. Sparrow. The march of the robot dogs. *Ethics and Information Technology*, 4(4):305–318, 2002.
- [93] R. Sparrow and L. Sparrow. In the hands of machines? The future of aged care. *Minds and Machines*, 16(2):141–161, 2006.
- [94] C.M. Stanton, P.H. Kahn Jr, R.L. Severson, J.H. Ruckert, and B.T. Gill. Robotic animals might aid in the social development of children with autism. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 271–278, 2008.
- [95] J.P. Sullins. When is a robot a moral agent? *International Review of Information Ethics*, 6(12):23–30, 2006.
- [96] Surgeon General’s Office. Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07, Final Report, Nov. 17, 2006.
- [97] A. Tapus, C. Tapus, and M.J. Mataric. Music therapist robot for individuals with cognitive impairments. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 297–298, 2009.
- [98] R. Tonkens. A challenge for machine ethics. *Minds and Machines*, 19(3):421–438, 2009.
- [99] US Department of Defense, Office of the Assistant Secretary of Defense (Public Affairs) News Transcript. DoD Press Briefing with Mr. Weatherington from the Pentagon Briefing Room, 18 Dec. 2007, [Online]. Available: <http://www.defense.gov/Transcripts/Transcripts.aspx?TranscriptID=4108>.
- [100] G. Veruggio. The birth of roboethics. Presented at *Workshop on Robo-Ethics, IEEE International Conference on Robotics and Automation (ICRA)*, 2005, [Online]. Available: <http://www.roboethics.org/icra2005/veruggio.pdf>
- [101] G. Veruggio. The euron roboethics roadmap. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pages 612–617, 2006.
- [102] G. Veruggio and F. Operto. Roboethics: a bottom-up interdisciplinary discourse in the field of applied ethics in robotics. *International Review of Information Ethics*, 6:2–8, 2006.
- [103] K. Wada and T. Shibata. Living with seal robots its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5):972–980, 2007.
- [104] K. Wada, T. Shibata, T. Musha, and S. Kimura. Robot therapy for elders affected by dementia. *IEEE Engineering in Medicine and Biology*, 27(4):53–60, 2008.
- [105] K. Wada, T. Shibata, T. Saito, K. Sakamoto, and K. Tanie. Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2785–2790, 2005.
- [106] M. Walzer. *Just and unjust wars*. Basic Books, 3rd edition, 2000.
- [107] V. Wiegel and J. van den Berg. Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, 1(3):233–242, 2009.
- [108] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow. KASPAR - A Minimally Expressive Humanoid Robot for Human-Robot Interaction Research. *Appl. Bionics and Biomech*, 6(3):369–397, 2009.